SEGMENT-BASED STEREO MATCHING*

By

Gerard G. Medioni and Ramakant Nevatia

Intelligent Systems Group
University of Southern California
Los Angeles, California 90089-0272

The [illegible]

Abstract

Images are two dimensional projections of three dimensional scenes, therefore depth recovery is a crucial problem in Image Understanding, with applications in passive navigation, cartography, surveillance, and industrial robotics. Stereo analysis provides a more direct quantitative depth evaluation than techniques such as shape from shading, and its being passive makes it more applicable than active range finding imagery by laser or radar. This paper addresses the subproblem of identifying corresponding points in the two images. The primitives we are using are groups of collinear connected edge points called segments, and we base the correspondence on the minimum "differential disparity" criterion. The result of this processing is a sparse array disparity map of the analyzed scene.

I. Introduction

The human visual system perceives depth with no apparent effort and very few mistakes, but how it does so is not understood. Binocular stereopsis plays a key role in this process, and the straightforward extraction of depth it provides, once corresponding points are identified, makes it very attractive. Depth recovery is necessary in domains such as passive navigation[Gennery80, Moravec80], cartography[Kelly77, Panton78], surveillance[Henderson79] and industrial robotics. Proposed solutions for the stereo problem follow a paradigm involving the following steps[Barnard82]:

-image acquisition,
-camera modeling,
-feature acquisition,
-image matching,
-depth determination,
-interpolation.

The hardest step is image matching, that is identifying corresponding points in two images, and this paper is solely devoted to it. The next section reviews the existing systems that have been proposed so far, divided in two broad classes, area-based and edge-based, then we summarize our assumptions and give a formal description of the method. The fourth section presents results, and we then discuss extensions.

II. Review of existing methods

Two classes of techniques have been used for stereo matching, area-based and feature-based.

2.1. Area-based stereo

Ideally, one would like to find a corresponding pixel for each pixel in each image of a stereo pair, but the semantic information conveyed by a single pixel is too low to resolve ambiguous matches, therefore we have to consider an area or neighborhood around each pixel, and use correlation-based matching algorithms to determine the corresponding match, it is therefore using local context to resolve ambiguities. The justification for such an approach is that of "continuity", that is disparity values change smoothly, except at a few depth discontinuities. All systems based on area-correlation suffer from the same limitations:

- They require the presence of a detectable texture within each correlation window, therefore they tend to fail in featureless or repetitive texture environments.
- They tend to be confused by the presence of a surface discontinuity in a correlation window.
- They are sensitive to absolute intensity, contrast and illumination.
- They get confused in rapidly changing depth fields (vegetation.)

For these reasons, the existing systems, specially the ones used in "automatic" cartography, require the intervention of human operators to guide them and correct them. Such systems are described in [Lucaa81, Panton78, Hannah80, Barnard80, Moravec79].

*This research was supported by the Defense Advanced Research Projects Agency and was monitored by the Air Force Wright Aeronautical Laboratories under Contract No. F33615-82-K-1786, DARPA Order No. 3119.

2.2. Feature-based systems

The depth information in stereo analysis is conveyed by the differences in the two images of a stereo pair due to the different viewpoints, the differences being most prominent at the discontinuities, or edges. Obviously, matching of features will not provide a full depth map, and must be followed by an interpolating scheme. The common characteristics of feature-based matching techniques are:

- They are faster than area-based methods, because there are many fewer points to consider.
- The obtained match is more accurate, edges can even be located with sub-pixel precision[Binford81].
- They are less sensitive to photometric variations, since they represent geometric properties of a scene.

Henderson[Henderson79] considered scenes representing cultural sites (man-made structures) and matched edge points on epipolar lines in the two views. He reduced ambiguity by assuming continuity between consecutive epipolar lines. Marr and Poggio have relied on two apparently simple constraints[Marr79]:

1. Uniqueness.
   Each point in an image may be assigned at most one disparity value. One may note that this assumption is not correct for transparent objects.
2. Continuity.
   Matter is cohesive, therefore values change smoothly, except at a few depth discontinuities.

They first proposed a cooperative algorithm[Marr76] that works very well on random-dot stereograms, but they rejected it to propose one of more heuristic nature, implemented by Grimson[Grimson79, Grimson81] that generates good results, given the very few assumptions. Arnold[Arnold78] mstches edges using local context, and his system seems to perform well on cultural scenes. Finally, Baker and Binford[Baker82] match edges on epipolar lines by using the no-reversal constrsint that the order of the match has to be preserved, in addition to uniqueness and continuity. They also consider continuity by examining adjacent epipolar lines. This system appears to perform reasonably on a wide variety of images.

In most of the systems presented above, a considerable saving in search time is obtained by a coarse to fine mstching, that is the msching is originally done on a low-resolution version of the imsge and the results are propsgated to the higher resolution version. However, it should be noted that in current implementations, good matches as well as errors tend to propagate from one level to the next.

III. The Minimal Differential Disparity Algorithm

From the survey conducted above, it appears that feature-based techniques are more appropriate to solve the correspondence problem, but edges as a primitive seem to be too low-level, and a connectivity check is needed to remove spurious matches. High level primitives such as physical object boundaries or surface descriptions would be preferred, however, stereo processing may need to precede the computstion of such descriptions. As a step towards higher level primitives, we are using segments. In order to generate them, we fit straight lines through adjacent edge points with a given tolerance of one pixel. These segments can be described by:

- coordinates of the end points
- orientation
- strength (average contrast)

By using these primitives, we implicitly assume the connectivity constraint. When matching segments, we need to allow one segment to possibly match with more than one segment in the other image (i.e. to allow for fragmented segments), even if we wish to preserve unique matches for the individual edge points. Also, instead of considering one epipolar line at a time, we have to consider all epipolar lines in which a given segment appears.

3.1. Assumptions and Definitions

We consider a simple camera geometry in which the epipolar plane, defined as the plane passing through an object point and the two camera foci, intersects the two image planes, so defining epipolar lines parallel to the y axis. Therefore, corresponding points must lie on corresponding epipolar lines, that is have the same row value, this is illustrated in Figure 3-1.

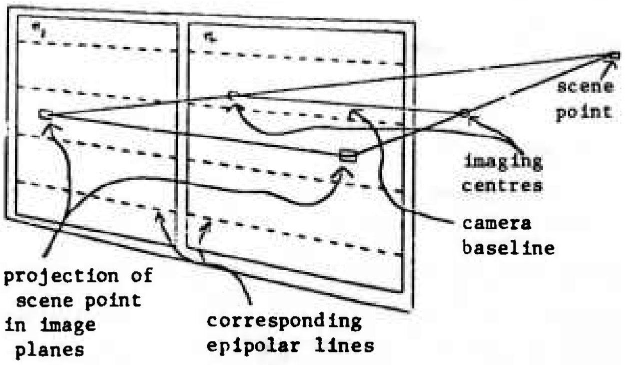
scene point
imaging centres
camera baseline
projection of scene point in image planes
corresponding epipolar lines

Figure 3-1: Collinear Epipolar Geometry from [Baker82]

We also give a bound on the disparity range allowable for any given segment, let us call it maxd.
Let $A=\{a_i\}$ be the set of segments in the left image
Let $B=\{b_j\}$ be the set of segments in the right image.
Then, for each segment a_i(resp. b_j) in the left (resp. right) image, we can define a <u>window</u> $w(i)$(resp. $w(j)$) in which corresponding segments from the right (resp. left) image must lie. The shape of this window is a parallelogram, one median being a_i(resp. b_j), the other a horizontal vector of length 2*maxd. One can see that a_i in $w(j)$ implies b_j in $w(i)$.
We define the boolean function $p(i,j)$ relating two segments as:

- $p(i,j)$ is true if
- b_j overlaps $w(i)$
- a_i , b_j have "similar" contrast
- a_i , b_j have "similar" orientation

The required similarity in orientation is loose and is a function of the segment length. We have set it to be 25 degrees for long segments and up to 90 degrees for very short segments.
Two segments are defined to have similar contrast if the absolute value of the difference of the individual contrasts is less than 20% of the larger one.
To each pair (i,j) such that $p(i,j)$ is true we associate an <u>average disparity</u> d_{ij} which is the average of the disparity between the two segments a_i and b_j along the length of their overlap.
We define the two functions S1 and S2 as:

$$S1(a_i)=\{j \mid b_j \text{ in } w(i) \text{ and } p(i,j) \text{ is true}\}$$
$$S2(a_i)=\{j \mid b_j \text{ in } w(i) \text{ and } p(i,j) \text{ is false}\}$$

Similarly, we define $S1(b_j)$ and $S2(b_j)$. We will also need the value $card(a_i)$, which is the number of elements in the set $S1(a_i)\ S2(a_i)$.
It is to be noted that all the functions described above are static, meaning that they are computed only once.

3.2. Description

Each possible match is evaluated by computing a measure of the distortion this match provokes for its neighbors,i.e. given that (i,j) is a correct match with its associated disparity d_{ij}, how well do the neighbors agree with this propoaed disparity? We compute an evaluation of the match (i,j) and compare to the matchea (i,k) and (h,j) for k in $S1(a_i)$ and h in $S1(b_j)$. If the evaluation is minimum for (i,j), then j is the preferred interpretation for i and i is the preferred interpretation for j. For any iteration after the first one, in order to evaluate a match (i,j), we only look at the preferred matches for the neighbors of i and j, if they have any. Formally, the computation of $v^t(i,j)$ is:

At iteration 1

$$v^1(i,j)=\left(\sum_{a_h \in S_1(b_j) \cup S_2(b_j)} \min_{\substack{b_k \in S_1(a_h) \\ b_k \neq h_j}} |d_{hk}-d_{ij}|\right)\Big/card(b_j)$$
$$+\left(\sum_{b_k \in S_1(a_i) \cup S_2(a_i)} \min_{\substack{a_h \in S_1(b_k) \\ a_h \neq a_i}} |d_{hk}-d_{ij}|\right)\Big/card(a_i)$$

At the end of each iteration, we define the cets $Q(a_i)$ and $Q(b_j)$ as

j in $Q(a_i)$ and i in $Q(b_j)$ if

$\forall k$ in $S1(a_i)$, $v^t(i,j) \leq v^t(i,k)$

AND

$\forall h$ in $S1(b_j)$, $v^t(i,j) \leq v^t(h,j)$

For any iteration after the first one, the computation of $v^t(i,j)$ becomes

$$v^t(i,j)=\left(\sum_{a_h \in S_1(b_j) \cup S_2(b_j)} \min_{\substack{b_k \in Q(a_h) \\ b_k \neq b_j}} |d_{hk}-d_{ij}|\right)\Big/card(b_j)$$
$$+\left(\sum_{b_k \in S_1(a_i) \cup S_2(a_i)} \min_{\substack{a_h \in Q(b_k) \\ a_h \neq a_i}} |d_{hk}-d_{ij}|\right)\Big/card(a_i)$$

if the sets Q are not empty, otherwise the computation of the function v is done using the formula for iteration 1.

At the last iteration, only those elements that have a preferred match are considered valid, and a disparity map array is filled using these values. It is interesting to note that this process is absolutely symmetric in the two views and therefore will yield identical results (except for the sign of the disparity) if the two views are interchanged. It is helpful to look at a simple example to understand this process.

3.3. Example

Let our 2 views be the ones shown in Figure 3-2 below:

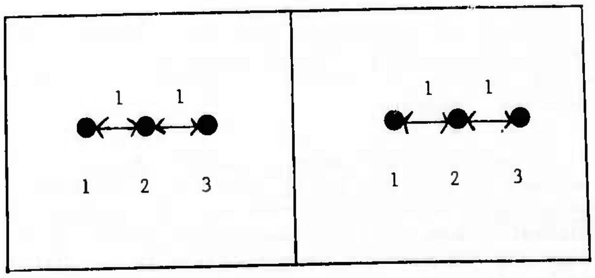

Figure 3-2: A aimple example

In absence of any extra information, the correct interpretation is that the 3 points have the same disparity, and the result of the matching is (a_i, b_i) for i in $\{1,2,3\}$.
In this example, $S1(a_i)=S1(b_i)=\{1,2,3\}$ and $S2(a_i)=S2(b_j)=\emptyset$. The array d_{ij} is

$$\begin{array}{rrr} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{array}$$

Therefore we find

$$\begin{aligned} v^1(1,1) &= (|d_{22}-d_{11}|+|d_{33}-d_{11}|)/3 \\ &+ (|d_{22}-d_{11}|+|d_{33}-d_{11}|)/3 \\ &= 0 \end{aligned}$$

compared to

$$\begin{aligned} v^1(1,2) &= (|d_{23}-d_{12}|+|d_{33}-d_{12}|)/3 \\ &+ (|d_{21}-d_{12}|+|d_{23}-d_{12}|)/3 \\ &= 1 \end{aligned}$$

and to

$$\begin{aligned} v^1(1,3) &= (|d_{22}-d_{13}|+|d_{32}-d_{13}|)/3 \\ &+ (|d_{12}-d_{13}|+|d_{11}-d_{13}|)/3 \\ &= 2.67 \end{aligned}$$

The calculations are similar for the other pairs, so, at the end of the first iteration, the preferred interpretations are only the correct ones, and further iterations will not alter the results.

3.4. Discussion

The criterion used here, namely the minimal differential disparity, has similarities with the edge interval constraints given in [Arnold80] and subsequently used by Baker[Baker 82], but looser in the sense that it does not require ordering of the edges. Since our criterion does not take ordering into account, a dynamic programming implementation is not possible. Our evaluation function is more informed than Baker's in the sense that it considers all edges in a neighborhood instead of just the predecessor and successor of a given edge. The performance of this algorithm on a few examples is presented next.

IV. Results

It is difficult to display results of stereo matching meaningfully, especially in a two dimensional picture, since we only generate a sparse disparity map. We will simply show the line segments in the two views that are found to match. We have not been able to master the art of cross-eyed stereo fusion, but since a number of people in the field are good at it, we will present all pairs of images according to its convention, that is the left view is shown on the right and the right view on the left. All results will also be shown this way, without explicitly marking each point and its correspondence. We first started our experiments with very simple line drawings, slightly more complex than the one shown in Figure 3-2 and the results matched the expectations. In order to remove the effects of the segmentation procedure on the performance of our matching technique, we hand-segmented the images shown in Figure 4-1 by tracing the boundaries of the objecta on a digitizing table. This image, from Control Data Corporation, is synthetic and has been used by Baker[Baker82] for his experiments. The resulting segments are shown on Figure 4-2 and Figure 4-3 displays the results after matching. All the lines that have been matched have the correct correspondence, but some matches are missed. This is due to the fact that when the matcher gets confused by closely competing assignments, it chooses not to assign a label. Also, some edges are not matches because of mistakes in the tracing procedure: we traced the boundaries of some objects in opposite directions in the two views.
For all other examples, edge detection was performed automatically using a technique developed by Nevatia and Babu[Nevatia80] that finds edge magnitude and direction by convolving the image with edge masks in different orientations (we used 5x5 masks in 6 directions here). These edges are then linked to form boundary curves which are approximated by piecewise linear segments.

Next, consider the industrial part shown in Figure 4-4, the original resolution is 256 by 256 and the gray levels are coded on 8 bits. We applied the matching algorithm to two different resolutions of the image, running it through three iterations. It was found that no assignment was changed after three iterations in our experiments. Figure 4-5 shows the original edges and Figure 4-6 displays the results in the above mentioned form. Similarly, Figure 4-7 shows the segments at half resolution and Figure 4-8 the results. Looking at the segments one by one, we did not notice any spurious assignment at either resolution, meaning that we captured the shape of the object, even though the density of edges is much larger than in the previous example.

Another, more complex image is shown on Figure 4-9. In this image, we have a wide range of disparities, a change of sign in the disparities across the picture, various occlusions, the presence of a repetitive structure (a Rubik's cube) and contrast reversal. We do not expect to get good results with this contrast reversal since one of our preliminary conditions is similarity in contrast, but the other peculiarities are very interesting. We worked at low resolution on the segments shown in Figure 4-10 to obtain the results shown in Figure 4-11. The interesting points are the following:

- The elongated vertical blocks in the rear of the image are correctly put into correspondence.

- All the squares of the cube that should be identified are correctly matched. The correct labeling appeared at iteration 2 (at iteration 1, most of them are only ambiguously matched.)

The segments at high resolution are shown in Figure 4-12 and the matching results in Figure 4-13. We did not use the results at low resolution to guide the matching at high resolution, therefore the elongated block in the rear right is not matched any longer. It is interesting to note that the edges coming from the texture of the wood blocks do not create confusion, but help the matching, on the front cylinder for example. Once again, most assigned matches are correct.

V. Conclusions

This research is far from being in a final state. The initial encouraging results presented here must therefore only be viewed as an indication that the hypothesis of minimal differential disparity may be useful. The critical points that must be examined are:

- Relax the contrast constraint. This may be done by considering not the contrast of an edge, but the intensity values on each side. Edges could then be matched if either their left side or their right side correspond. One may eventually consider an edge as a doublet[Baker82] and match each side separately.
- To refine the formulation of the evaluation formula. Statistical analysis may yield better functions, maybe by introducing a static probability measure to evaluate each match based on similarity of intrinsic properties (length, color, orientation.) Also of concern is a more accurate definition of a no-match label, which is obtained if a match pair is not clearly better than the competing ones.
- Further extensive testing is also required on aerial and near range imagery, with terrain models for accuracy checking.
- Finally, we must use an interpolation scheme, very likely intensity-based, to generate a full disparity map of the scene depth.

VI. References

Arnold78 Arnold D. "Local Context in Matching Edges for Stereo Vision," in Proceedings of Image Understanding Workshop, Cambridge, Mass, May 1978, pp. 65-72.

Arnold80 Arnold D. and Binford T. "Geometric constraints in Stereo Vision," Society Photo-Optical Instr. Engineers, Vol. 238, Image Processing for Missile Guidance, 1980, pp. 281-292.

Baker82 Baker H. "Depth from Edge and Intensity Based Stereo," Stanford Artificial Intelligence Laboratory, AIM 347, Stanford, Calif., Sept. 82.

Barnard80 Barnard S. and Thompson W. "Disparity Analysis of Images," IEEE Trans. Pattern Anal. Machine Intell., PAMI-2,4 July 1980, pp. 333-340.

Barnard82 Barnard S. and Fishler M. "Computational Stereo," ACM Computing Surveys, Vol. 14, No. 4, Dec. 1982, pp. 553-572.

Binford81 MacVicar-Whelan P. and Binford T. "Line Finding with Subpixel Precision," in Proceedings of Image Understanding Workshop, Washington, D.C., Apr. 1981, pp. 26-31.

Gennery80 Gennery D. "Object Detection and Measurement Using Stereo Vision," in Proceedings of Image Understanding Workshop, College Park, Md., Apr. 1980, pp. 161-167.

Grimson79 Grimson W. and Marr D. "A Computer Implementation of a Theory of Human Stereo Vision," in Proceedings of Image Understanding Workshop, Palo Alto, Calif., Apr. 1979, pp. 41-47.

Grimson81 Grimson W. "From Images to Surfaces," MIT Press, Cambridge, Mass., 1981.

Hannah80 Hannah M. "Bootstrap Stereo," in Proceedings of Image Understanding Workshop, College Park, Md., Apr. 1980, pp. 201-208.

Henderson79 Henderson R., Miller R. and Grosch C. "Automatic Stereo Reconstruction of Man-Made Targets," SPIE, Vol. 186, No. 6, Digital Processing of Aerial Images, 1979, pp. 240-248.

Kelly77 Kelly R. McConnell P. and Mildenberger S. "The Gestalt Photomapping System," Journal of Photogrammetric Engineering and Remote Sensing, Vol. 43, No. 1407, 1977.

Lucas81 Lucas B. and Kanade T. "An Iterative Image Registration Technique with an Application to Stereo Vision," in Proceedings of Image Understanding Workshop, Washington, D.C., Apr. 1981, pp. 121-130.

Marr76 Marr D. and Poggio T. "Cooperative Computation of Stereo Disparity," Science 194, 1976, pp. 283-287.

Marr77 Marr D. and Poggio T. "A Theory of Human Stereo Vision," Memo. 451, Artificial Intelligence Laboratory, MIT, Cambridge, Mass., Nov. 1977.

Moravec80 Moravec H. "Obstacle Avoidance and Navigation in the Real World by a Seeing Robot Rover," Stanford Artificial Intelligence Laboratory, AIM 340, Ph.D. Thesis, Sept. 1980.

Nevatia80 R. Nevatia and K. Babu, "Linear Feature Extraction and Description," Computer Graphics and Image Processing, Vol. 13, pp. 257-269, July 1980.

Panton78 Panton D. "A Flexible Approach to Digital Stereo Mapping," Journal of Photogrammetric Engineering and Remote Sensing, Vol. 44, No. 12, Dec. 1978, pp. 1499-1512.

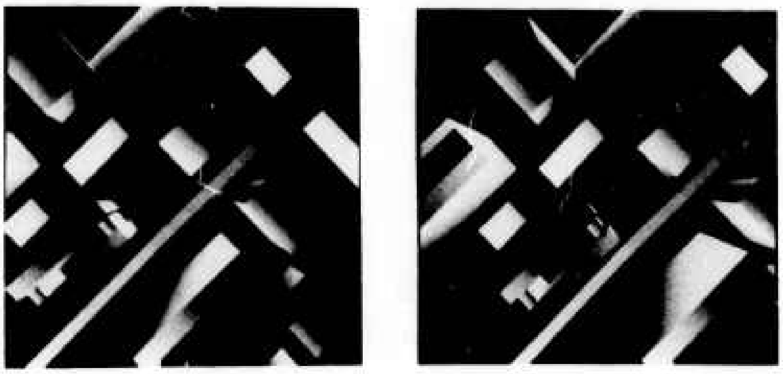

Figure 4-1: Synthetic image [256x256x6]

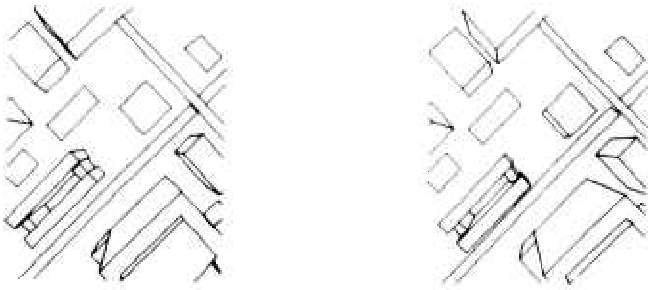

Figure 4-2: Hand generated segments

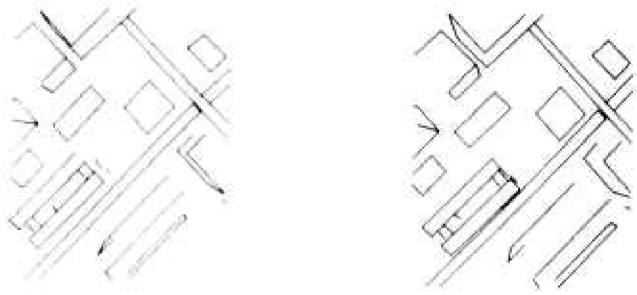

Figure 4-3: Results of the matching

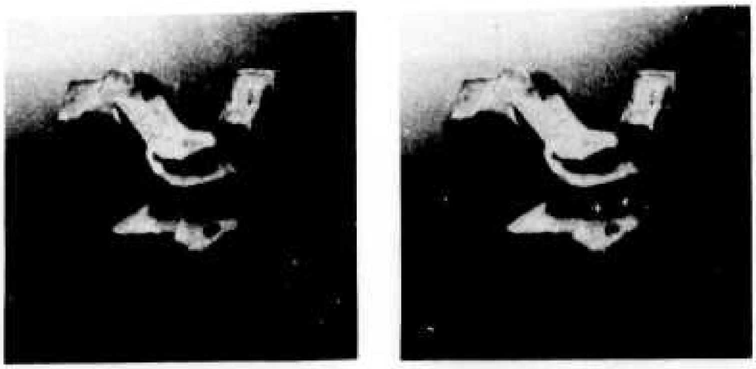

Figure 4-4: Industrial part [256x256x8]

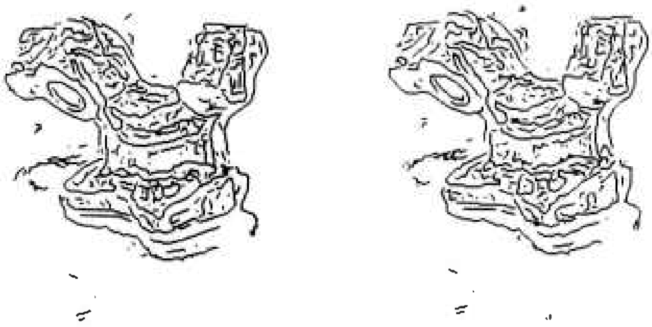

Figure 4-5: Segments from the full resolution image

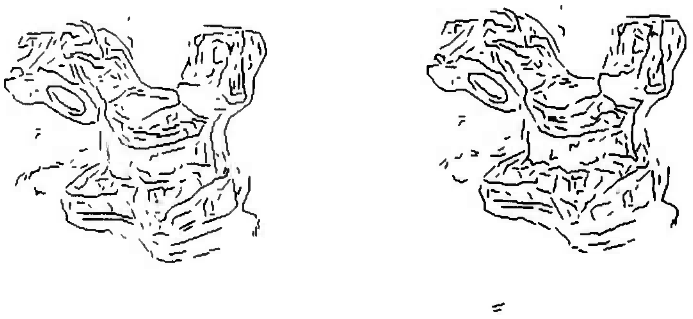

Figure 4-6: Results at full resolution

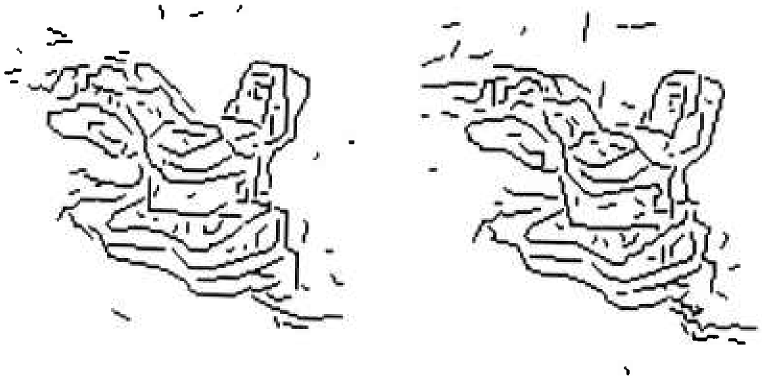

Figure 4-7: Segments from the half resolution image

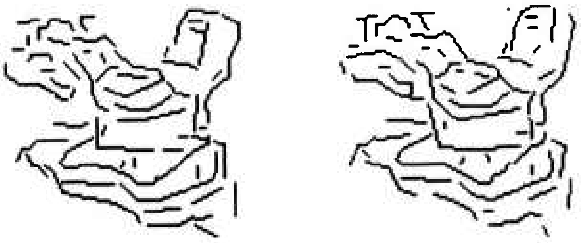

Figure 4-8: Results at half resolution

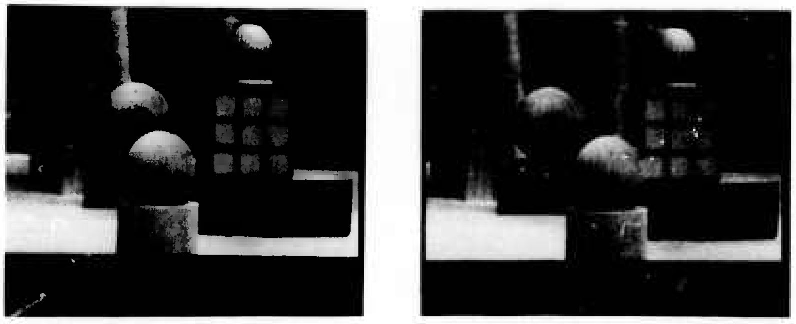

Figure 4-9: Image of some blocks[512x512x7]

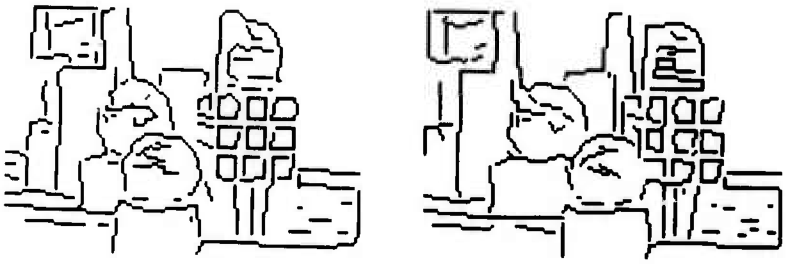

Figure 4-10: Segments at low resolution

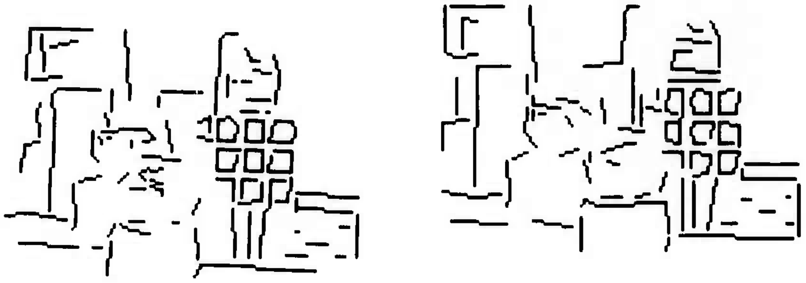

Figure 4-11: Results at low resolution

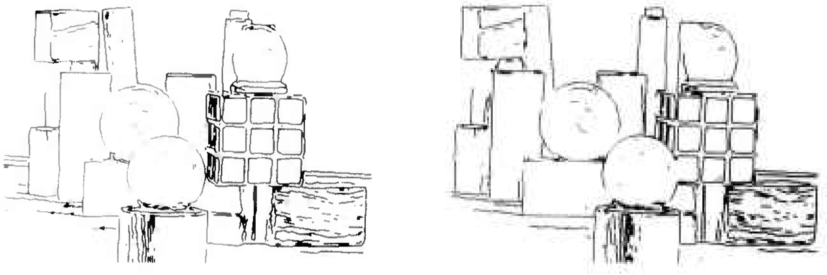

Figure 4-12: Segments at high resolution

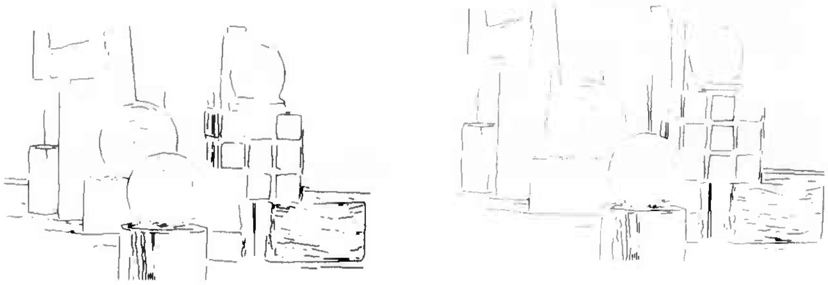

Figure 4-13: Results at high resolution